LC-005

仮想一元化 NAS システム X-NAS の同期バックアップ実現に向けた順序制御方式の検討

Consideration of Synchronized Backup of X-NAS: a Clustered NAS System

江端 淳† 川本 真一† 保田 淑子† 沖津 潤† 樋口 達雄† 濱中 直樹†
Atsushi Ebata† Shinichi Kawamoto† Yoshiko Yasuda† Jun Okitsu† Tatsuo Higuchi† Naoki Hamanaka†

1. はじめに

我々は，安価，使い勝手の良さ，高い容量拡張性という 3 つの特長を持つ仮想一元化 NAS システム X-NAS(eXpandable NAS)を提案している[1]。これまで，NAS 上のファイルが障害によって消失することを防ぐため，NAS 上のファイルをテープ装置にバックアップする方法が一般的に用いられてきた。しかし，テープ装置は高価であり，また，バックアップ時間も長い。そこで，テープ装置より安価で高速な NAS にバックアップする NAS to NAS バックアップ[2]が用いられるようになってきた。しかし，NAS to NAS バックアップは非同期的に行われるため，最新ファイルがバックアップされていない状況でバックアップ元に障害が発生すると，最新ファイルが消失してしまう。

そこで，X-NAS では，X-NAS 上のファイルが障害によって消失するのを防ぐため，NAS をバックアップ先とした同期バックアップ(NAS to NAS 同期バックアップ)をサポートする。X-NAS は，市販の NAS を要素 NAS として使用できるように，業界標準のリモートファイルシステムプロトコルである NFS を用いて仮想一元化を実現している[1]。この考え方を NAS to NAS 同期バックアップに応用し，バックアップのための専用機能を持たない一般の NAS をバックアップ NAS として使用できるようにする。同期バックアップは，X-NAS とバックアップ NAS の両方に同時に NFS プロシージャを発行することにより実現する。NFS プロシージャは必ずしもクライアントが発行した順に処理されるわけではない。このため，X-NAS とバックアップ NAS との間で完全な同期を実現するには，NFS プロシージャの順序制御を実現する必要がある。本稿では X-NAS の NAS to NAS 同期バックアップの実現に向けた NFS プロシージャの順序制御方式について述べる。

2. X-NAS の概要

X-NAS のシステム構成を図 1 に示す。X-NAS は複数の要素 NAS からなる NAS クラスタである。ファイルは要素 NAS のいずれか 1 つに格納される。要素 NAS の 1 つを親 NAS，それ以外を子 NAS と呼ぶ。親 NAS はエントリ NAS に Xnfsd と管理ディスクを加えたものである。Xnfsd は複数の要素 NAS を仮想一元化するデーモンである。管理ディスクはファイルがどの要素 NAS に格納されているかを記録する。子 NAS は X-NAS のための専用機能を持たない市販のエントリ NAS である。Xnfsd は UNIX クライアントから NFS プロシージャを受けると，管理ディスクを参照してアクセス対象ファイルの格納先要素 NAS を特定する。そして，特定した要素 NAS にそのプロシージャを転送し，NFS デーモン(nfsd)に処理させる。一方，Windows クライアントから CIFS リクエストを受けると，Samba がそれを NFS プロシージャに変換し，Xnfsd が変換結果を処理する。

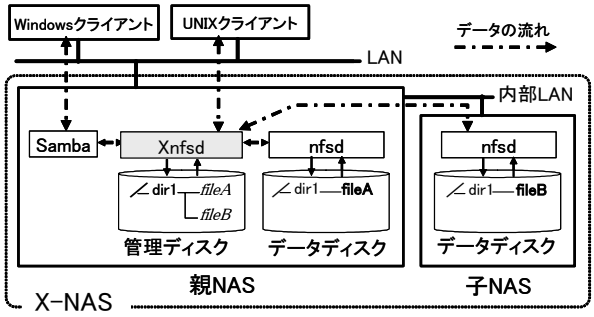

図 1: X-NAS のシステム構成

3. NAS to NAS 同期バックアップ

図 2 に X-NAS の NAS to NAS 同期バックアップ方式を示す。バックアップ NAS は NFS サーバの機能を持つ NAS であればどんな NAS でもよく，X-NAS であっても良い。X-NAS の Xnfsd には同期バックアップ機能を追加する。Xnfsd が書込み系プロシージャ[1]を受けると，それを要素 NAS とバックアップ NAS の両方に書き込み，同期バックアップを実現する。例えば，Xnfsd が NFS クライアントから WRITE プロシージャを受けると(①)，Xnfsd は管理ディスクにアクセスし(②)，ファイルの格納先要素 NAS を特定し，受けた WRITE プロシージャをバックアップ NAS と格納先要素 NAS の両方に転送する(③④)。両 NAS の nfsd は，Xnfsd が転送した WRITE プロシージャを受け，データをディスクに書込み，Xnfsd に結果を返す(⑤⑥)。Xnfsd は，バックアップ NAS と格納先要素 NAS から結果を受けると，クライアントに結果を返す(⑦)。他の書込み系プロシージャも同様に処理される。上記処理により，NFS プロシージャレベルで X-NAS とバックアップ NAS のファイルを常に一致させる同期バックアップを実現する。

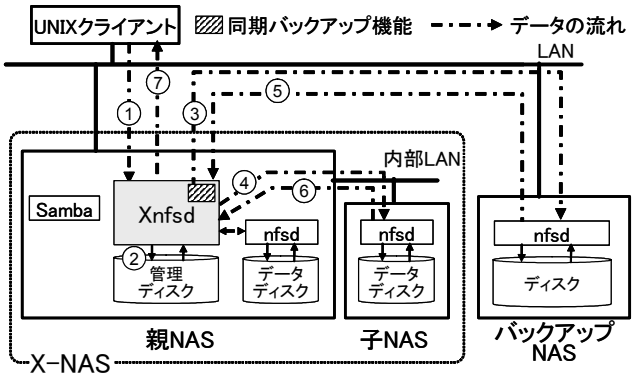

図 2: X-NAS の NAS to NAS 同期バックアップ方式

4. NFS プロシージャの順序制御方式

4.1. 順序制御の必要性

前章で説明したように，Xnfsd はクライアントから書込み系 NFS プロシージャを受けると，X-NAS の要素 NAS とバ

†(株)日立製作所中央研究所

[1] WRITE，CREATE，RENAME，REMOVE，SETATTR 等

ックアップ NAS の両方に，そのプロシージャを転送する。クライアント側でロックをとれば，X-NASとバックアップNASにおけるプロシージャ処理の順序を保障できるが，プログラマのミス等によりクライアント側ではロックをとらない場合も考えられる。このようなケースでは，要素 NAS とバックアップNAS間のプロシージャの処理順序が入れ替わる可能性がある。プロシージャ処理の順序が入れ替わると両 NAS において異なるデータが書き込まれてしまう。その結果，要素 NAS とバックアップ NAS 上のデータの不一致が発生し，同期バックアップを実現できない。

例えば，Xnfsd が異なる 2 つのクライアントからほぼ同時に WRITE プロシージャ Wa，Wb を受信した場合を考える。Wa，Wb は同一オフセットにそれぞれデータ Da，Db を書込むものとする。Xnfsd は Wa を X-NAS の要素 NAS とバックアップ NAS の両方に転送する。このとき，バックアップ NAS に転送した Wa が途中で消失したとする。Xnfsd はバックアップ NAS に対して Wa を再送するが，その前に Wb を受信したとする。このとき，要素 NAS は Wa，Wb の順にプロシージャを処理し，バックアップ NAS は Wb，Wa の順にプロシージャを処理する。その結果，要素 NAS には Db が，バックアップ NAS には Da が書込まれてしまう。

この問題を解決するため，Xnfsd において書込み系プロシージャの処理順序を保障する必要がある。

4.2. 順序制御方式

X-NAS がクライアントから受けた先行プロシージャと後続プロシージャの間で順序制御が必要な場合は，以下の 4 つに分けられる。

(C1) 先行プロシージャと後続プロシージャが同一ファイルに書込む場合。前述したように，X-NAS とバックアップ NAS でプロシージャの処理順序が異なると，両者の間にデータの不一致が発生する。

(C2) 先行プロシージャがディレクトリの作成や属性変更を行い，後続プロシージャがそのディレクトリを含む下位ディレクトリにファイル又はディレクトリを書込む場合。たとえば，先行の SETATTR がディレクトリの属性をアクセス禁止に変更し，後続の CREATE がそのディレクトリ直下にファイルを生成すると，後続の CREATE はエラーとなる。しかし，処理順序が逆転すると CREATE はエラーとならずファイルが生成される。このため，X-NAS とバックアップ NAS の間にファイルの不一致が発生する。

(C3) 先行プロシージャがファイル又はディレクトリを作成し，後続プロシージャがその上位のディレクトリの属性を変更する場合。たとえば，先行の CREATE がファイルを生成し，後続の SETATTR がそのファイルの上位のディレクトリをアクセス禁止にする場合，X-NAS とバックアップ NAS のどちらか一方で処理順序が逆転すると，逆転した NAS の CREATE 処理でエラーが発生し，X-NAS とバックアップ NAS の間にデータの不一致が発生する。

(C4) 先行プロシージャがファイル又はディレクトリを削除又は移動し，後続プロシージャがその親ディレクトリを削除する場合。たとえば先行の REMOVE がファイルを削除し，後続の RMDIR がそのファイルの親ディレクトリを削除する場合，X-NAS とバックアップ NAS のどちらか一方で処理順序が逆転すると，逆転した NAS の RMDIR の処理でエラーが発生し，X-NAS とバックアップ NAS の間にディレクトリの不一致が発生する。

上記の 4 つをまとめたものが表 1 である。*f* は順序制御不要の場合である。*f* の例として先行がファイルの REMOVE，後続が WRITE の場合，まず管理ディスクで対象ファイルのエントリが削除され，後続の WRITE は管理ディスクアクセス時にエラーとなりデータディスクにアクセスしないため，順序制御不要となる。Xnfsd はクライアントから受けた先行プロシージャと後続プロシージャの種別を表 1 に照らし合わせる。その結果が *C1*～*C4* であれば，図 3 の定義に従って真偽を判定し，真であれば順序制御必要，偽であれば順序制御不要と判定する。Xnfsd は，クライアントから書込み系プロシージャを受けると，先行して処理中の書込み系プロシージャがなければ，そのままそのプロシージャを処理する。先行プロシージャがあれば，表 1 と図 3 を用いて順序制御の必要性を判定し，必要である場合には先行プロシージャの処理が完了するまで後続プロシージャの処理を待たせる。

表 1: 順序制御判定表

先行プロシージャ		後続プロシージャ	WRITE	CREATE	RENAME	REMOVE	SETATTR	MKDIR	RENAME	RMDIR	SETATTR
ファイル		WRITE	C1					f	C3	f	C3
		CREATE						C1		C4	
		RENAME									
		REMOVE	f					f		f	
		SETATTR									
ディレクトリ		MKDIR	f				f	C2	C2∥C3	C1	C3
		RENAME	C2							C3	C2∥C3
		RMDIR	f	C1			f	C1	C3	C4	C3
		SETATTR	C2					C2	C2∥C3	C2	C2∥C3

・先行プロシージャのアクセス対象のパス名 Pp の定義
Pp := /NP*1*/NP*2*/…/NP*n*
・後続プロシージャのアクセス対象のパス名 Pf の定義
Pf := /NF*1*/NF*2*/…/NF*m*
・C1～C4 の定義
C1 := (Pp = Pf)　{ 先行 Pp と後続 Pf が同一であれば真 }
C2 := (*n*≦*m*) and (NP*1* = NF*1*) and … and (NP*n* = NF*n*)
{ 先行 Pp が後続 Pf より上位であれば真 }
C3 := (*n*≧*m*) and (NP*1* = NF*1*) and … and (NP*m* = NF*m*)
{ 先行 Pp が後続 Pf より下位であれば真 }
C4 := (*n*+*1* = *m*) and (NP*1* = NF*1*) and … and (NP*n* = NF*n*)
{ 先行 Pp が後続 Pf の直下であれば真 }

図 3: 条件 C1～C4 の定義

5. おわりに

本稿では仮想一元化 NAS システム X-NAS の NAS to NAS 同期バックアップの実現に向け，課題となるNFSプロシージャの順序制御方式について検討した。本方式により，クライアントプログラムのミス等により発生しうる，X-NAS の要素 NAS とバックアップ NAS において NFS プロシージャの処理順序が入れ替わる問題を解決でき，NFS プロシージャを用いた同期バックアップを実現できる。

参考文献

[1] 川本真一，他，“ファイル自律配置方式を備えた仮想一元化 NAS システム X-NAS の実現と評価”，DEWS2003，4-B-01，2003.

[2] Microsoft Corporation，“Deploying Windows Powered NAS Using Dfs with or Without Active Directory”,http://www.microsoft.com/windows/storage/productinformation/whitepapers，2001.